

* Read Employee and make a Lookup
* Redirect the “No Matching” to a Fuzzy Lookup
* Union both and use a OLE DB Destination

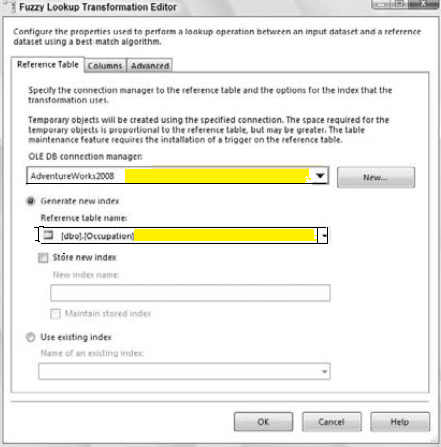
The transform will create several output columns that you may or may not decide are useful to store in a table. Either way they are important to understand:

**Input and Pass - Through Field Names and Values** — This column contains the name and value of the text input provided to the Fuzzy Lookup Transform or passed through during the lookup

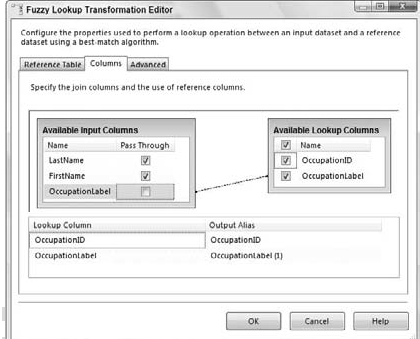
**Reference Field Name and Value** — This column contains the name and value(s) of the matched results from the reference table

**Similarity** — This column contains a number between 0 and 1 representing similarity. Similarity is a threshold that you set when configuring the Fuzzy Lookup Transform. The closer this number is to 1, the closer the two text fields must match

**Confidence** — This column contains a number between 0 and 1 representing confidence of the match relative to the set of matched results. Confidence is different from similarity, because it is not calculated by examining just one word against another but rather by comparing the chosen word match against all the other possible matches. Confidence gets better the more accurately your reference data represents your subject domain, and it can change based on the sample of the data coming into the ETL process



Using the Fuzzy Lookup Transform requires at least one field to be a string, either a DT\_WSTR or DT\_STR data type. On the Columns tab in the editor you need to map at least one text field from the input to the reference table for comparison

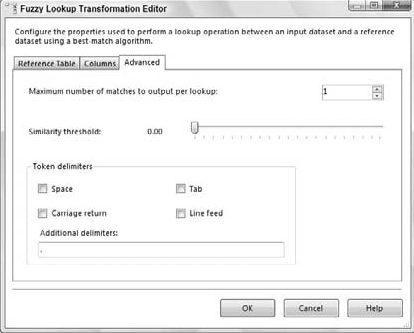


The Advanced tab contains the settings that control the fuzzy logic algorithms.

You can set the maximum number of matches to output per incoming row.

The **default is set to 1,** which means pull the best record out of the reference table if it meets the similarity threshold. Incrementing this setting higher than the default might generate more results that you’ll have to sift through, but it might be required if there are too many closely matching strings in your data.

A **slider controls the similarity** threshold. When you are experimenting, a good strategy is to start this setting at **0.5** and move up or down as you review the results. This setting is normally decided based on a businessperson’s review of the data, not the developer’s review. If a row cannot be found that’s similar enough, the columns that you checked in the Columns tab will be set to NULL. The token delimiters can also be set if, for example, you don’t want the comparison process to break up incoming strings by a period (.) or spaces.



SETTING “UNION ALL”

